MegaProto/E: Power-Aware High-Performance Cluster with Commodity Technology

Taisuke Boku* Mitsuhisa Sato* Daisuke Takahashi* Hiroshi Nakashima†
Hiroshi Nakamura‡ Satoshi Matsuoka§ Yoshihiko Hotta*

* University of Tsukuba, {taisuke,msato,daisuke,hotta}@hpcs.cs.tsukuba.ac.jp
† Toyohashi University of Technology, nakasima@tutics.tut.ac.jp
‡ University of Tokyo, nakamura@hal.rcast.u-tokyo.ac.jp
§ Tokyo Institute of Technology, matsu@is.titech.ac.jp

Abstract

In our research project named "Mega-Scale Computing Based on Low-Power Technology and Workload Modeling", we have been developing a prototype cluster by using not ASIC or FPGA but only commodity technology. Its packaging is extremely compact and dense, and its performance/power ratio is very high. Our previous prototype system named "MegaProto" demonstrated that one cluster unit which consists of 16 commodity low-power processors can be successfully implemented on just 1U height chassis and it achieves up to 2.8 times higher performance/power ratio than ordinary high-performance dual-Xeon 1U server unit.

We have improved MegaProto by replacing the CPU and enhancing the I/O performance. The new cluster unit named "MegaProto/E" with 16 of Transmeta Efficeon processors achieves 32GFlops of peak performance which is 2.2-fold of the original one. The cluster unit is equipped with independent dual network of Gigabit Ethernet including dual 24-port switches. The maximum power consumption of the cluster unit is kept to 320W which is comparable with today's high-end PC server for high performance clusters.

Performance evaluation using NPB kernels and HPL shows that the performance of MegaProto/E exceeds that of dual-Xeon server in all the benchmarks, and its performance ratio ranges from 1.3 to 3.69. These results reveal that our solution of implementing a number of ultra low-power processors in compact packaging is an excellent way to achieve extremely high performance on applications with certain degree of parallelism. We are now building a multi-unit cluster with 128 CPUs (8 units) to prove that this advantage still holds with higher scalability.

1. Introduction

The advantage of PC cluster solution for high performance computing has been continued in recent years with its high performance/cost ratio supported by commodity technology. However, there is an apprehension for its applicability to highly scalable systems with tens of thousands or millions of processors. In order to clear the apprehension, we at least have to solve the problem of power consumption, space and dependability. Our research project named "Mega-Scale Computing Based on Low-Power Technology and Workload Modeling" aims to establish fundamental technologies for this purpose. Our research covers the feasibility and dependability of the million-scale parallel systems as well as the programmability on such a large scalability.

For the feasibility study focusing on the performance/power/space ratio problem especially, we have been developing the prototype system based on commodity-only technology, that is, we only use commodity processors and network elements. To realize a high density implementation, however, we develop cluster chassis units which can contain a number of processors in a small space. As shown in recent trend of dual-core CPUs such as PentiumD or OpteronD, it is clear that the best way to improve the total performance is to introduce multi-processor with relatively low-performance instead of increasing the CPU clock frequency on a single processor. Based on this concept, one of the ideal platform is a cluster with ultra low-power processors implemented on a small space with high density.

Green Destiny[1] is a successful example of the above concept. While it consists of commercial blade-style processor card, we have designed and implemented more dense collection of processors in 1U hight chassis with 17 of Transmeta Crusoe processors. This prototype unit was named "MegaProto"[2, 6]. The cluster unit consists

of 16 computation nodes and one management node, each of them is equipped with 933 MHz Crusoe processor, 256 MByte SDRAM and dual-port Gigabit Ethernet NIC attached through a 32bit/33MHz PCI bus. Since its processor supports single floating point operation per clock, the total peak performance of a cluster unit is 14.9 GFlops. In this version of prototype, the network performance was bottlenecked by the poor PCI bus which we had to equip due to the power budget limitation.

When we designed the first prototype of MegaProto, we also planned the enhanced version with more powerful processor and I/O bus, and now it was completed to build the second version named "MegaProto/E" ('E' stands for Efficeon, the name of new processor). In this paper, we describe the design, implementation and performance evaluation of MegaProto/E cluster unit.

The rest of the paper consists of the followings. Section 2 gives the outline of our Mega-Scale computing project and the conceptual design of MegaProto series cluster unit. In Section 3, we describe the detailed design and implementation of MegaProto/E. After describing the performance evaluation in Section 4, the cluster design with higher scalability based on MegaProto/E is shown in Section 5. Finally, the conclusions and future works are described in Section 6.

2. Mega-Scale Project and MegaProto Cluster Unit

The overview of our Mega-Scale Computing Project and the conceptual design were basically described in [2]. In this section, we describe them briefly as the minimum knowledge to read this paper.

2.1. Overview of Mega-Scale Computing Project

Today, the Peta-Flops computing is not just a dream anymore, and several projects have been launched aiming this class of computational performance. In these projects, the common key issues include (i) how to implement ultra large-scale systems, (ii) how to reduce the power consumption per Flops, and (iii) how to control ultra large-scale parallelism. From the viewpoint of hardware technology, the first two issues are essential. As shown by BlueGene/L[3], the state-of-the-art MPP today, one of the promised way to the Peta-Flops computing is to build an MPP with very low-power processors and simple switch-less network to save both the space and power consumption. Such a system is, however, realized by a dedicated hardware platform including specially designed processor chips, network and system racks, which require a large amount of system cost as well as the long time for design and implementation.

On the other hand, the rapid progress of computation and communication technologies not limited for HPC applications lead us to another solution based on low-power commodity technology both on the CPU and network. While an ordinary PC cluster for HPC applications consists of high-performance CPUs consuming tens of Watts and wide-bandwidth network such as Infiniband[4], we can aggregate a number of very low-power and medium-speed CPUs aiming laptops and Gigabit Ethernet with very high performance/cost supported by non-HPC market.

In our Mega-Scale Computing Project[5], we researches (i) hardware/software cooperative low-power technology and (ii) workload modeling and model-base management of large scale parallel tasks as well as the faults occurring in the system. Our research covers the processor architecture, compiler, networking, cluster management and programming based on the above concept.

2.2. Conceptual Design of MegaProto

In this section, we briefly introduce the conceptual design of MegaProto. The word of "MegaProto" is used for our overall prototyping system for the feasibility study on our Mega-Scale Computing concept. However, we called our first version of prototype with the same name in the past literatures[2, 6]. To distinguish two versions of prototype systems, we explicitly call the first one and second one as "MegaProto/C" and "MegaProto/E", respectively. Two suffix words 'C' and 'E' stand for Crusoe and Efficeon, respectively, the code name of CPUs to be used.

When we consider the Peta-Flops scale system with commodity CPU and network technologies, the issues on power consumption and spacing are not avoidable. As a rough sketch, we consider the boundary conditions of them as 10MW for power consumption and 1,000 racks for space, which are still hard but possible. From these conditions, our primary goal is to achieve 1TFlops/10kW/rack as the performance/power/space ratio. MegaProto is designed as a series of prototype systems for the verification of it.

Assuming the commodity system formation, we implement the system with 19-inch 42U height of standard racks. Except the space for network switch, 32U can be assigned for computation nodes. Thus, the goal can be converted to realizing 32GFlops/310W/1U building block as the basic unit. It cannot be achieved with today's high-end CPUs such as Intel Xeon, Intel Itanium2 or AMD Opteron, in either or both on performance and power consumption even with multi-way SMP configuration. On the other hand, state-of-the-art low power CPUs with DVS and very low voltage drive are possible to achieve it if we can aggregate 10 to 20 of CPUs in 1U chassis. Even with blade-style cards, this density is impossible. However, we finally solve the problem to develop a mother board to contain 16 of daughter cards with one-dollar-bill size.

Another important issue on the unit chassis is the interconnection network suitable for such a basic design. If we adopt a high-end CPU in the node processor, we also

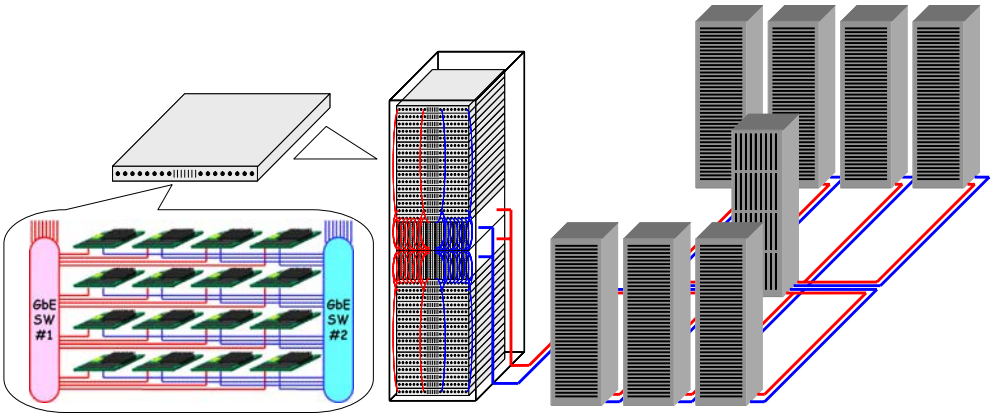


Figure 1. System configuration

need a high-end network interface with 1GByte/sec bandwidth or more for efficient parallel processing. Since we consider the mid-range CPU as the node processor, it can be reduced to several hundreds of MByte/sec as the suitable network bandwidth. This range can be covered by trunking of several channels of Gigabit Ethernet, and we concluded to use dual channels per processing node. It is quite important from the view point of performance/cost ratio to introduce commodity Gigabit Ethernet as the interconnection. Since the recent Gigabit Ethernet switching fabric is quite inexpensive and small for 10 to 20 ports of connection, it is possible to implement the basic switch itself on the mother board to connect all processing nodes which are mounted on the same mother board. Hereafter, we call this 1U chassis of the building unit containing multiple CPUs and intra-connection network switches as "cluster unit".

Figure 1 shows the conceptual view of the 1U cluster unit and the overall system. In the figure, 16 of CPUs which are equipped with two channels of Gigabit Ethernet NICs and individual two switches are mounted on a cluster unit. A system rack contains 32 cluster units and interconnection switches, and finally hundreds of system racks realizes a Peta-Flops system.

3. Design and Implementation of MegaProto/E

In this section, we describe the detailed design and implementation of MegaProto/E compared with the previous version, MegaProto/C[2, 6].

3.1. Implementation of MegaProto/C

Before introducing the detailed design and implementation of MegaProto/E, first we show the implementation of the previous model. We planned to develop the MegaProto as the first and second versions according to the availability of the parts and modules. When we started the design and implementation of MegaProto, Transmeta Crusoe was the best candidate as the CPU to be used, and IBM Japan had already provided a processor card with CPU, memory and I/O bus extension as a commercial product for embedded controlling system. In MegaProto/C, Transmeta Crusoe (TM5800) with 933MHz clock frequency was employed of which peak performance was 933MFlops because it can issue only a single floating point operation per clock. Thus, the peak performance with 16 CPUs on a cluster unit is limited to 14.9 GFlops, and it could not reach to our goal described in the previous section.

However, we considered it as a good start for the development because we just had to develop the mother board to contain 17 of these processor cards as daughter cards. Therefore, we decided to develop the mother board at the first stage of the plan and to develop a new daughter card later with further available Efficeon processor of which production schedule fitted to our two-staged plan. Actually, MegaProto/C (with Crusoe) played the roll of "prototype of prototype" for software development including Linux kernel tuning, drivers for NICs and switches, compiler and MPI library settings as well as fixing the environment of power consumption measuring[7].

On the cluster unit, there are two categories of interconnection networks, the data network and the management network. Hereafter, 16 of processor cards for computation are called "computation nodes" while a processor card for system management is called "management node".

Data Network: It consists of two individual Gigabit Ethernet with switches. Each processor card is equipped with dual Gigabit Ethernet ports, and each port is connected to a 24-port Gigabit Ethernet switch (Broadcom BCM5692). Since only the computation nodes are connected to this network, there are unconnected 8 ports on each switch. These 8 links are connected to external RJ-45 ports for inter-unit connection outside the clus-

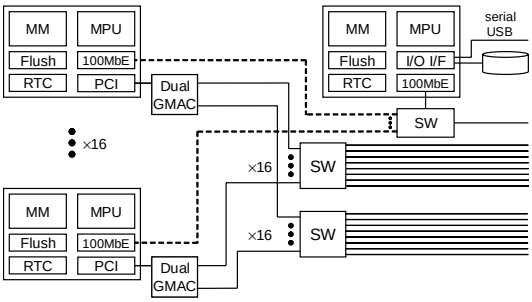


Figure 2. Block diagram of MegaProto/C cluster unit

ter unit (See Section 5). A computation node can drive both network links simultaneously for trunking (double bandwidth) or duplicated communication (fault tolerance) by software. This network is mainly used for data exchanging on parallel processing.

Management Network: It consists of a Fast Ethernet with a switch. All computation nodes and management node are equipped with a Fast Ethernet port, and these links are bound to a Fast Ethernet switch with dual upper-level Gigabit Ethernet links (Broadcom BCM5646). These upper-level links are connected to external RJ-45 ports for inter-unit connection. This network is mainly used for network management on the operating system (Linux) for NIS, NFS, remote login, remote shell, etc.

Figures 2 and 3 show the block diagram and the real picture of MegaProto/C, respectively. Only the management node is equipped with a 2.5inch hard disk drive with 60GByte capacity to contain all system files for 17 nodes in the cluster unit. At the system boot time, all disk-less computation nodes are booted via the Management Network sharing the binary images on the HDD of management node. User's home directories are build on the outside file server to be shared by multiple cluster units through the external links of the Management Network.

3.2. Design of MegaProto/E

As described in the previous subsection, we designed and implemented the second version of MegaProto while constructing the software environment of MegaProto/C and evaluating it. The new version of cluster unit is named "MegaProto/E" (with Efficeon). On this version, we developed a new processor card (daughter card) equipped with enhanced processor, memory and I/O bus from MegaProto/C. Several minor changes were also performed on the mother board to improve the system stability.

The processor card was designed to fit the connection socket of the mother board of MegaProto/C, however the

Figure 3. Picture of MegaProto/C cluster unit

density of each processor card became higher than that of MegaProto/C. The processor card consists of two small PCBs which are vertically stacked. The I/O bus to connect the processor card to the Data Network was improved from 32bit/33MHz PCI to 64bit/66MHz PCI-X, which provides the four times bandwidth of the old one. It provides 533MByte/sec of theoretical peak bandwidth to support dual bisection Gigabit Ethernet links of which peak bandwidth is 500MByte/sec. Due to employing the commercial embedded controller module as the computation node on MegaProto/C, there was a severe bandwidth bottleneck on it. It was improved by this upgraded I/O bus and reflected to several benchmark performance (See Section 4). The memory throughput was also improved from SDR-133 to DDR-266 as well as doubled capacity on MegaProto/E.

Since the computation performance is not directly limited by the performance of the management node, the processor card for the management node was kept as the same as MegaProto/C, that is, we did not use Efficeon processor here. Therefore, the management node and computation nodes have a heterogeneous CPU configuration on MegaProto/E. There is no actual problem at this point.

3.3. Implementation of MegaProto/E

As described above, the main work for the implementation of MegaProto/E was performed to the processor card for computation nodes. The improvements on the processor card from that of MegaProto/C is summarized in Table 1. Especially, the enhancements on memory throughput and PCI bus are expected to be reflected to the performance improvements both on the single CPU performance and the parallel processing performance derived by high network bandwidth.

Although the TDP of each CPU is reduced to less than

	MegaProto/C	MegaProto/E
MPU	TM5800 (0.93GHz)	TM8820 (1.0GHz)
TDP	7.5W	3W
Peak Perf./Power	124.0 MFlops/W	666.7 MFlops/W
Caches	L1=64KB(I)+64KB(D)	L1=128KB(I)+64KB(D)
	L2=512KB(D)	L2=1MB(D)
Memory	256 MB SDR (133 MHz)	512 MB DDR (266 MHz)
Flush	512 KB	1 MB
I/O Bus	PCI (32 bit, 33 MHz)	PCI-X (64 bit, 66 MHz)

Table 1. Processor card specification

the half of Crusoe's one, the power consumption on the memory module and the bus and the PCI-X bridge are increased. As a result, the power consumption of each processor card is slightly increased, and the total power consumption of MegaProto/E cluster unit is 320W at maximum while that of MegaProto/C is 300W. However, this small fraction of power increase is acceptable for largely enhanced memory and I/O bus performance as well as more than twice of floating point performance on the CPU. The real picture of the cluster unit is shown in Figure 4.

The most important performance improvement on MegaProto/E is that we can obtain excellent performance/power/space ratio, that is, 1.024TFlops/10.24kW/rack which satisfies our goal.

4. Performance Evaluation

In this section, we evaluate the basic performance of a single cluster unit of MegaProto/E comparing with that of MegaProto/C and ordinary high-end PC server with dual Xeon in SMP configuration. For the viewpoint of the network performance comparison, we also refer the performance of two-node system with the same configuration of dual-Xeon servers.

The benchmark programs referred in this evaluation are commonly used ones, HPL (High Performance Linpack)[8] and NPB (Nas Parallel Benchmarks)[9] kernels. For NPB kernels, the problem size is class-A. For HPL, the performance with $N = 10,000$ is shown. All sources are compiled with gcc/g77 version 3.2.2, linked with LAM-MPI version 7.1.1, and executed under Linux kernel version 2.4.22mmpu. The environment of dual-Xeon 1U server with similar software environments but slightly different versions; gcc/g77 version 3.4.3, LAM-MPI version 6.5.6 and Linux kernel 2.4.20-20.7smp. On all benchmarks, only a single channel Gigabit Ethernet is used to avoid software overhead of trunking and to keep the fairness with two-node dual-Xeon servers.

Figure 4. Picture of MegaProto/E cluster unit

4.1. Performance Improvements from MegaProto/C

Figure 5 shows the overall performance comparison between MegaProto/C and MegaProto/E. In all graphs, the speed-up ratios to the performance of MegaProto/C with 4 CPUs are shown.

As shown in these results, MegaProto/E obtains from 1.06 to 2.38 times of performance gain from MegaProto/C. We analyzed these results as follows:

FT, MG: 1.37 $\sim$ 1.65 times of performance gain is obtained basically by the improvement of floating point operation speed which is more than twice of MegaProto/C.

CG: The communication data amount in CG is larger than other benchmarks. The performance improvement on PCI-X bus provides 2.38 times of performance gain.

HPL: 1.74 times of performance gain is obtained by the upgraded floating point performance as like as FT and MG. However, the efficiency of Linpack performance to the theoretical peak is only 30.1% due to the small capacity of memory. MegaProto/C achieved 38% of peak performance with 16 processors, and the memory capacity problem is serious on MegaProto/E with powerful floating point performance.

IS: Almost no performance gain because the lack of floating point operations. The gain of CPU clock frequency is only 7%, thus it is reasonable.

EP: The performance result seems to be strange because EP is basically a floating point bound benchmark. We guess that this benchmark implies a large number of fundamental numerical functions such as *log* or *sqrt* which may not be well-tuned on ordinary gcc math li-

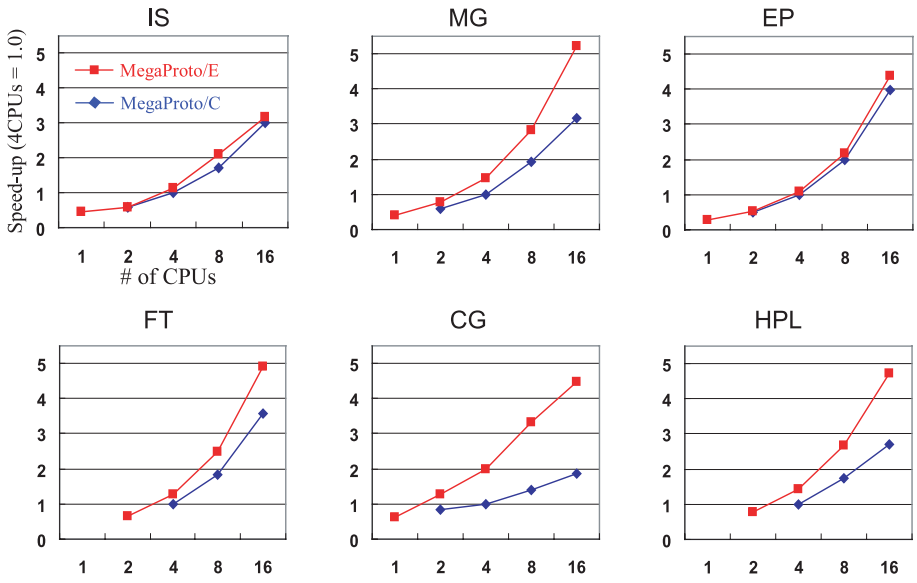


Figure 5. Speed-up comparison between two versions of MegaProto

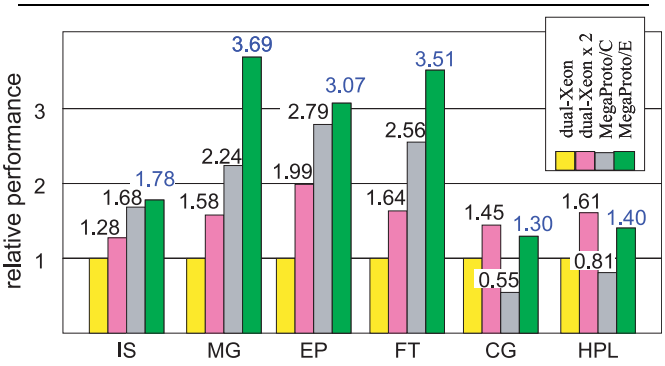


Figure 6. Comparison with Xeon-base servers

brary. Then the large overhead of function calls inhibits the performance gain.

In total, we can see the good performance gain especially for memory and network throughput bound benchmarks to be reasonably improved. The memory capacity problem on HPL is serious, and it shows that our solution is not suitable for non-fine grained applications.

4.2. Comparison with Xeon-base Systems

Figure 6 shows the relative performance among dual-Xeon, two-node dual-Xeon, MegaProto/C and MegaProto/E. Dual-Xeon system is a single node PC server in SMP configuration, and two-node dual-Xeon (labeled as "dual-Xeon x 2") is a small cluster to connect two of dual-Xeon nodes with a single Gigabit Ethernet. All performance are shown in the relative one based on the performance of dual-Xeon. The PC server used here is Appro 1124Xi with 3.06 GHz Intel Xeon and 1GByte of DDR memory. Since the total TDP and peak performance of processors, 170W and 12.2GFlops respectively[10], and its maximum AC power rating of the entire 1U system, 400W, are all comparable to our cluster unit with 16 processors, the dual-Xeon server is a good reference for performance comparison.

At first, we can see the performance of MegaProto/E always exceeds that of dual-Xeon ranging from 1.30 to 3.69 times of improvement. In MG, EP and FT, it achieves remarkable score. Especially for MG and FT, which are CPU performance and memory throughput bound benchmarks, MegaProto/E achieves the excellent performance. Although MegaProto/C marked less performance than dual-Xeon for CG and HPL, MegaProto/E overcomes it with the improved network bandwidth and increased memory capacity.

Except CG and HPL, MegaProto/E marks higher score than two-node dual-Xeon system although our system runs with less than the half power consumption of two-node dual-Xeon. Since both systems are based on a single channel Gigabit Ethernet, it means that MegaProto/E is equipped an interconnection network with better performance balance between CPU performance and network bandwidth than Xeon-base system. With this scenario, we can estimate the scalability of MegaProto solution is much better than Xeon-base HPC cluster if we only adopt commodity Ethernet as the interconnection network. Such a performance balance is quite important for large scale parallel processing systems, and it is shown that our solution based on com-

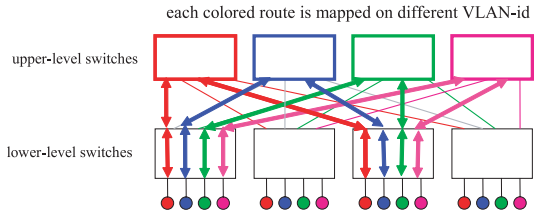


Figure 7. Fat Tree Network with VLAN-based routing

modity technology imported from non-HPC world works well in this world.

5. Multi-Unit System

5.1. How to Utilize Multiple Upper-Links

After the confirmation of the excellent performance of MegaProto/E, we are building multi-unit system with more than 16 processors. At the first stage, we will build a cluster with 128 processors connecting 8 of MegaProto/E cluster units. Since MegaProto/E cluster unit is equipped with 8 ports of external links for each Data Network on a channel, the total potential bandwidth from the cluster unit is 16 Gbps or 2GByte/sec with dual channel of Data Network.

However, there is a problem to utilize 8 uplink ports for external inter-unit connection. Since the on-board Gigabit Ethernet switches on MegaProto/E is in Layer-2, so-called "broadcast storm" occurs if we simply connect two or more uplink ports of multiple cluster units. The broadcast storm occurs when multiple links make more than one loops including the node PC and any of intermediate Layer-2 switches.

There are two ways to solve this problem:

1. Using Layer-3 switches which are featured the IP-base routing function, and connect all intermediate links through these switches. All looped connections are logically cut and all loops disappears.
2. Using tagged-VLAN (Virtual LAN) for pseudo static routing to separate multiple links in isolated domain, and cutting the loops as shown in Figure 7[11].

The first method is simple but it requires expensive Layer-3 switches to support a large number of Gigabit Ethernet ports. Such a switch costs more than US$10,000 and this method is not acceptable in ordinary situation. The second method is reasonable because most of today's medium class of Layer-2 switches are equipped with tagged-VLAN (IEEE802.1q)[12] feature. However, it requires multiple IP addresses on each node, and very complicated and tricky IP route setting is necessary on the whole network with standard VLAN driver on Linux[11].

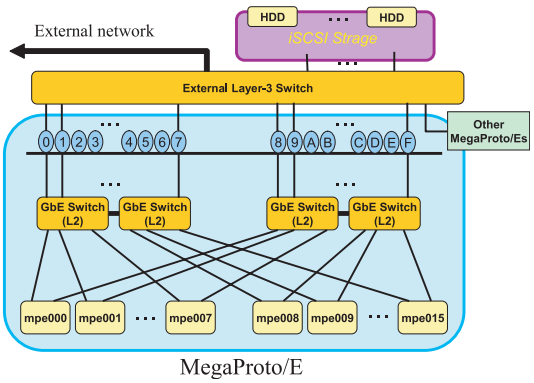


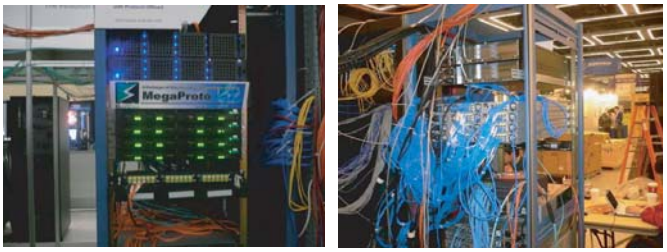
Figure 8. Block diagram of StorCloud with MegaProto/E at SC2005

To solve the problem in sophisticated way, we developed a special device driver to handle the source-destination IP routing with attaching/detaching VLAN-tag on Ethernet frames. With this technique, we can build a flat IP-space network with single IP address per node to exploit the multiplied bandwidth on multiple upper-level links and switches. This network system is called VFREC-NET (VLAN-based Flexible, Reliable and Expandable Commodity Network)[13]. For example, we need 8 sets of Layer-2 8-port Gigabit Ethernet switches with IEEE802.1q feature as the upper-level switches to combine 8 of MegaProto/E cluster units. Then we configure the whole network where each uplink from a cluster node is connected to one of these 8 switches. On the upper-level switches, from any source to any destination cluster unit, there exists a unique link which can be tagged as one of 4095 tags which is the physical limit of IEEE802.1q protocol.

The on-board Gigabit Ethernet switch on MegaProto can handle this mechanism, and we can scale-up the entire system to utilize all the allowed tags. It seems expensive to introduce 8 of 8-port switches, but actually a switch can be virtualized into multiple logical switches with VLAN, and if we carefully select the assigned VLAN-tag without any conflict in the system, it is possible to configure the system network with a minimum set of Layer-2 switches.

5.2. MegaProto/E Multi-Unit System at SC2005 StorCloud Challenge

In some special cases, we can utilize the first method described above. For StorCloud demonstration in SC2005 at Seattle, we brought four sets of MegaProto/E cluster unit and operated them as iSCSI server with 64 processors through 64 Gigabit Ethernet links. The operation of StorCloud challenge was performed by AIST (Advanced Institute of Science and Technology, Japan)[14], and these MegaProto/E units ran through the event with just a sin-



Front View Rear View (with 64 cables)

Figure 9. StorCloud challenge with MegaProto/E at SC2005

gle hardware failure on a Gigabit Ethernet port of a computation node. However, this failure could be tolerated using the other port of Data Network, and the total bandwidth was not lost.

Figures 8 and 9 show the block diagram of MegaProto connection to the iSCSI server through Layer-3 switch and the picture of MegaProto/E at StorCloud corner at SC2005, respectively. Through the demonstration, we could confirm the stability of MegaProto/E cluster unit and its applicability for variety of usage not limited to the computational cluster computing.

6. Conclusions and Future Works

We have been developing a building block of large-scale power-aware PC cluster for high performance computing based on only commodity processor and network technologies, named MegaProto. Its latest version named MegaProto/E cluster unit with Transmeta Efficeon processor achieves 32GFlops/320W/1U of performance/power/space density suitable for mounting on standard 19-inch rack. Including the space for inter-unit network switches, we can construct 1TFlops/10kW/rack Linux ready cluster based on dual-link Gigabit Ethernet.

The benchmark results show that our MegaProto/E demonstrates much better performance on NPB kernels and HPL than ordinary high-end PC server with dual Intel Xeon in SMP configuration, keeping the same space occupancy and less power consumption. It is proved that typical applications with certain degree of parallelism can be effectively solved on our platform which leads us to very high density scalable cluster systems.

Not only building the hardware platform, we are also developing the software tools to connect these cluster units for up to thousands of processors in a system based on commodity Gigabit Ethernet routing with VLAN technology.

Our future work includes the construction of a medium size system with hundreds of processors, performance evaluation of the system including the network performance equipped with our VLAN solution, the verification of fault tolerance software not only for the processing node but also for the interconnection network.

Through the research on Mega-scale computing based on low-power technology and workload modeling, we will continue to seek the effective way towards the Peta-Flops computing environment.

Acknowledgments

The authors would express their appreciation to technical staff of IBM Japan for their contributions and support. This research work is supported by Japan Science and Technology Agency as a CREST research program entitled “Mega-Scale Computing Based on Low-Power Technology and Workload Modeling.”

References

[1] M. Warren, et al., “High-density computing: A 240-node Beowulf in one cubic meter”, in Proc. Supercomputing 2002, Nov. 2002.

[2] H. Nakashima, et al., “MegaProto: A Low-Power and Compact Cluster for High-Performance Computing”, in Proc. HPPAC05 (with IPDPS2005), Apr. 2005.

[3] N. R. Adiga et al., “An overview of the BlueGene/L supercomputer”, in Proc. Supercomputing 2002, Nov. 2002.

[4] http://www.infiniband.org/

[5] Mega-Scale research team, “Mega-Scale computing based on low-power technology and workload modeling”, http://www.para.tutics.tut.ac.jp/megascale/, 2005.

[6] H. Nakashima, et al., “MegaProto: 1 TFlops/10kW Rack Is Feasible Even with Only Commodity Technology”, in Proc. Supercomputing 2005, Nov. 2005.

[7] Y. Hotta, et al., “Measurement and characterization of power consumption of microprocessors for power-aware cluster”, in Proc. COOL Chips VII, Apr. 2004.

[8] A. Petitet, et al., “HPL - a portable implementation of the high-performance Linpack benchmark for distributed-memory computers”, http://www.netlib.org/benchmark/hpl/, Jan. 2004.

[9] D. H. Bailey et al., “The NAS parallel benchmarks”, in Proc. Intl. J. Supercomputer Applications, 5(3):63-73, 1991.

[10] Intel Corp. Datasheets of the following Intel processors on 90nm process: Xeon (302355-001), Pentium 4 (303128-004), Mobile Pentium 4 (302424-002), Celeron M (300302-003) and Pentium M (302189-004), 2004.

[11] T. Kudoh, et al., “VLAN-based Routing: Multi-path L2 Ethernet network for HPC Clusters”, in Proc. of CLUSTER2004, Sep. 2004.

[12] http://www.ieee802.org/1/pages/802.1Q.html

[13] S. Miura, et al., “Low-cost high-bandwidth tree network for PC clusters based on tagged-VLAN technology”, in Proc. ISPAN2005, Dec. 2005.

[14] O. Tatebe, et al., “High-performance KEKB/Belle data analysis using Gfarm Grid file system”, StorCloud Challenge, SC2005, Nov. 2005.